Statistical Mechanics of Min-Max Problems

Anonymous authors
Paper under double-blind review

Abstract

Min-max optimization problems, also known as saddle point problems, have attracted significant attention due to their applications in various fields, such as fair beamforming, generative adversarial networks (GANs), and adversarial learning. However, understanding the properties of these min-max problems has remained a substantial challenge. This study introduces a statistical mechanical formalism for analyzing the equilibrium values of min-max problems in the high-dimensional limit, while appropriately addressing the order of operations for min and max. As a first step, we apply this formalism to bilinear min-max games and simple GANs, deriving the relationship between the amount of training data and generalization error and indicating the optimal ratio of fake to real data for effective learning. This formalism provides a groundwork for a deeper theoretical analysis of the equilibrium properties in various machine learning methods based on min-max problems and encourages the development of new algorithms and architectures.

1 Introduction

Min-max optimization problems, also known as saddle point problems, are well-known classical optimization problems extensively studied in the context of zero-sum games (Wald, 1945; Von Neumann & Morgenstern, 2007). These problems have diverse applications across various fields, such as game theory, machine learning, and signal processing. In game theory, min-max problems arise in zero-sum games where one player's gain corresponds to another's loss. Several methods have been proposed to find the min-max value or equilibrium points in these games (Dem'yanov & Pevnyi, 1972; Maistroskii, 1977; Bruck, 1977; Lions, 1978; Nemirovsky et al., 1983; Freund & Schapire, 1999). In machine learning, min-max games are relevant for training generative adversarial networks (GANs) (Goodfellow et al., 2020; Arjovsky et al., 2017), where a generator creates synthetic fake data, and a discriminator distinguishes between fake and real data in a zero-sum game setting. Additionally, in adversarial learning, these problems are employed to train models that are robust to adversarial attacks by optimizing a worst-case perturbed loss function (Szegedy et al., 2013; Goodfellow et al., 2014b; Papernot et al., 2016; Madry et al., 2017),

Despite their widespread application of min-max optimization problems, several challenges still need to be addressed, including understanding the usefulness of these min-max formulations, evaluating the convergence properties of the algorithms, and conducting sensitivity analyses of min-max values. One promising approach to addressing these issues is to focus on the typical-case behavior of min-max problems in randomized instance ensembles, such as dataset ensembles. Statistical-mechanical approaches, which have demonstrated their effectiveness in analyzing the typical-case behavior of randomized instance ensembles of optimization and constraint-satisfaction problems (Mézard & Parisi, 1986; Fontanari, 1995), provide a powerful formalism for such analyses. Extending this formalism to analyze the typical-case behavior of min-max values thus presents a potential direction for further research, although this has not yet been fully explored. Therefore, this study applies the statistical mechanical formalism to min-max problems, modeling them as a virtual two-temperature system. This formalism enables sensitivity analysis of the typical-case min-max values in the high-dimensional limit. Notably, this formalism properly addresses the order of min-max operations, which is critical in non-convex scenarios where interchanging the order of min and max can lead to incorrect results. By using this formalism, we analyze typical-case min-max values of bilinear min-max games and simple GANs. In particular, we derive the relationship between the amount of training data and generalization error and indicate the optimal ratio of fake data to real data for effective learning.

Our main contributions are as follows:

- We introduce a statistical-mechanical formalism developed for sensitivity analysis of equilibrium values in high-dimensional min-max problems.
- Applying this approach, we conduct a detailed sensitivity analysis on a bilinear min-max game to verify the theoretical validity of our approach.
- Building on this formalism, we analyze the generalization performance of GANs and determine the optimal ratio between fake and real data for effective training.

Notation Here, we summarize the notations used in this study. We use the shorthand expression $[N] = \{1, 2, \ldots, N\}$, where $N \in \mathbb{N}$. $I_d \in \mathbb{R}^{d \times d}$ denotes a $d \times d$ identity matrix, and $\mathbf{1}_d$ denotes the vector $(1, \ldots, 1)^\top \in \mathbb{R}^d$ and $\mathbf{0}_d$ denotes the vector $(0, \ldots, 0)^\top \in \mathbb{R}^d$. For the matrix $A = (A_{ij}) \in \mathbb{R}^{d \times k}$ and a vector $\boldsymbol{a} = (a_i) \in \mathbb{R}^d$, we use the shorthand expressions $dA \triangleq \prod_{i=1}^{d} \prod_{j=1}^{k} dA_{ij}$ and $d\boldsymbol{a} \triangleq \prod_{i=1}^{d} da_i$, respectively.

2 Statistical Physics Formalism for Min-Max Optimization Problems

The section introduces a statistical-mechanical formalism that models min-max problems as a virtual two-temperature system from a statistical mechanics perspective. Min-max problems are formally expressed as

$$\Omega = \min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{y} \in \mathcal{Y}} V(\boldsymbol{x}, \boldsymbol{y}; A), \quad \text{s.t.} \ \ \boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}, \quad \boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y} \tag{1}$$

where $V(\cdot, \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is a bivariate function; $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$ are the optimization variables; $\mathcal{X}$ and $\mathcal{Y}$ are the feasible sets; A is a parameter characterizing the problem, e.g., graph G. We introduce the following Boltzmann distribution to analyze min-max problems for a given bivariate function $V(\boldsymbol{x}, \boldsymbol{y}; A)$ in Eq. (1), with virtual inverse temperatures $\beta_{\min}$ and $\beta_{\max}$:

$$p_{\beta_{\min}, \beta_{\max}}(\boldsymbol{x}; A) = \frac{1}{Z(\beta_{\min}, \beta_{\max}, A)} e^{-\beta_{\min}\left(\frac{1}{\beta_{\max}} \ln \int d\boldsymbol{y} e^{\beta_{\max} V(\boldsymbol{x}, \boldsymbol{y}; A)}\right)}, \tag{2}$$

where $Z(\beta_{\min}, \beta_{\max}, A)$ is the normalization constant, also known as the partition function. Hereafter, we refer to it as the partition function. By first taking the limit $\beta_{\max} \to \infty$ followed by $\beta_{\min} \to \infty$, the distribution $\lim_{\beta_{\min} \to +\infty} \lim_{\beta_{\max} \to +\infty} p_{\beta_{\min}, \beta_{\max}}(\boldsymbol{x}; A)$ concentrates on a uniform distribution over the min-max values. Note that the order of these limits is crucial because the min and max operations cannot be interchanged in non-convex and non-concave min-max problems. While a similar formulation has been used in previous work (Varga, 1998), they simultaneously take the limits of both $\beta_{\min}$ and $\beta_{\max}$ with a fixed ratio $\beta_{\min}/\beta_{\max} = \mathcal{O}((d_x)^0, (d_y)^0)$, which does not fully capture the distinct effects of the min and max operations in non-convex settings. Such an approach generally does not yield accurate results when the function $V(\boldsymbol{x}, \boldsymbol{y}; A)$ is non-convex to the variables.

Statistical-mechanical approaches have demonstrated their effectiveness in analyzing the typical-case behavior of randomized instance ensembles of optimization and constraint-satisfaction problems (Mézard & Parisi, 1986; Fontanari, 1995). This work also focuses on evaluating the typical cases of min-max problems characterized by a random parameter A. Our main objective is to calculate the logarithm of $Z(\beta_{\min}, \beta_{\max}, A)$ averaged over the random variables A in the limit $\beta_{\max} \to \infty$ followed by $\beta_{\min} \to \infty$:

$$\Omega = \lim_{\beta_{\min} \to \infty} \lim_{\beta_{\max} \to \infty} f(\beta_{\min}, \beta_{\max}) \tag{3}$$

where

$$f(\beta_{\min}, \beta_{\max}) = -\frac{1}{\beta_{\min} d_x} \mathbb{E}_A \log Z(\beta_{\min}, \beta_{\max}, A), \tag{4}$$

referred to as the free energy density. Setting the ratio of the inverse temperatures as $p = -{}^{\beta_{\min}}/_{\beta_{\max}}$, this can be rewritten as

$$
\begin{aligned}
f(\beta_{\min}, \beta_{\max}) &= -\frac{1}{\beta_{\min} d_x} \mathbb{E}_A \log \int d\boldsymbol{x} e^{-\beta_{\min}\left(\frac{1}{\beta_{\max}} \ln \int d\boldsymbol{y} e^{\beta_{\max} V(\boldsymbol{x}, \boldsymbol{y}; A)}\right)}, \\
&= -\frac{1}{\beta_{\min} d_x} \mathbb{E}_A \log \int d\boldsymbol{x} \left(\int d\boldsymbol{y} e^{\beta_{\max} V(\boldsymbol{x}, \boldsymbol{y}; A)} \right)^p . \qquad (5)
\end{aligned}
$$

Although calculating the expectation value of the logarithm is generally difficult, the replica method provides a resolution:

$$
\Omega = -\lim_{\beta_{\min}\to\infty} \lim_{\beta_{\max}\to\infty} \lim_{\gamma\to+0} \frac{1}{\beta_{\min} d_x \gamma} \log \mathbb{E}_A \left(\int d\boldsymbol{x} \left(\int d\boldsymbol{y} e^{\beta_{\max} V(\boldsymbol{x}, \boldsymbol{y}); A} \right)^p \right)^\gamma . \qquad (6)
$$

Instead of directly handling the logarithmic expression in Eq. (5), one calculates the average of the γ-th and p-th power for $\gamma, p \in \mathbb{N}$, performs an analytic continuation to $\gamma, p \in \mathbb{R}$ for this expression, and finally takes the limits $\gamma \to +0$, $\beta_{\min} \to +\infty$ and $\beta_{\max} \to +\infty$. Based on this replica "trick", the calculation can be simplified to the replicated partition function $Z_\gamma(\beta_{\min}, \beta_{\max})$:

$$
Z_\gamma(\beta_{\min}, \beta_{\max}) = \mathbb{E}_A \left(\int d\boldsymbol{x} \left(\int d\boldsymbol{y} e^{\beta_{\max} V(\boldsymbol{x}, \boldsymbol{y})} \right)^p \right)^\gamma = \mathbb{E}_A \int \prod_{a=1}^{\gamma} d\boldsymbol{x}^a \prod_{l=1}^{p} d\boldsymbol{y}^{al} e^{\beta_{\max} \sum_{a,l} V(\boldsymbol{x}^a, \boldsymbol{y}^{al}; A)}, \qquad (7)
$$

up to the first order of γ to take the $\gamma \to +0$ limit in the right hand side of Eq. (6). This computation is a standard procedure in the statistical physics of random systems and is generally accepted as exact, although rigorous proof has not yet been provided.

Additionally, before taking the limits, $\beta_{\min} \to \infty$ and $\beta_{\max} \to \infty$, the concepts of finite inverse temperature $\beta_{\min}$ and $\beta_{\max}$ correspond to scenarios where neither the minimum nor the maximum is fully achieved, a common situation in the min-max algorithm. This approach provides valuable insights into cases where neither extreme is fully realized or both are only partially optimized. Exploring novel algorithms based on this finite-temperature generalization of min-max problems represents an intriguing direction for future work. Furthermore, in game theory, this formalism can be interpreted as modeling games with relaxed assumptions of complete rationality.

In the following sections, we apply this formalism to a fundamental and significant bilinear min-max game, demonstrating that the analytic continuation of p is a rigorous operation. We then analyze the minimal model of GANs as a more practical example.

3 Bilinear Min-max Games

This min-max formalism introduces two replica parameters: γ, associated with the randomness of A, and p, related to the dual structure of min-max problems. The analytic continuation with respect to the replica parameter γ is widely recognized as effective and is frequently employed in the statistical mechanics of optimization. However, the analytic continuation of the replica parameter p has not been explored. While establishing its mathematical validity presents challenges, this study eliminates the influence of the replica parameter γ associated with the randomness of A and rigorously demonstrates that the analytic continuation with respect to p holds for fundamental bilinear min-max games. Specifically, we show that the free energy density derived using the replica trick in Eq. (6), as explained in Section 2, is equivalent to the exact expression in Eq. (5) derived without analytic continuation of γ and p for bilinear min-max games (Tseng, 1995; Daskalakis et al., 2017).

Bilinear games are regarded as a fundamental example for studying new min-max optimization algorithms and techniques (Daskalakis et al., 2017; Gidel et al., 2019; 2018; Liang & Stokes, 2019). Mathematically, bilinear zero-sum games can be formulated as the following min-max problem:

$$
\min_{\boldsymbol{x} \in \{0,1\}^{d_x}} \max_{\boldsymbol{y} \in \{0,1\}^{d_y}} V(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{W}), \qquad (8)
$$

where $V(\cdot,\cdot)$ is given by

$$V(\boldsymbol{x},\boldsymbol{y};\boldsymbol{W})=\frac{1}{2d_x}\boldsymbol{x}^\top W_{xx}\boldsymbol{x}+\frac{1}{2d_y}\boldsymbol{y}^\top W_{yy}\boldsymbol{y}+\frac{1}{\sqrt{d_xd_y}}\boldsymbol{x}^\top W_{xy}\boldsymbol{y}+\boldsymbol{x}^\top\boldsymbol{b}_x+\boldsymbol{y}^\top\boldsymbol{b}_y, \tag{9}$$

where $\boldsymbol{W}=(W_{xx},W_{yy},W_{xy},\boldsymbol{b}_x,\boldsymbol{b}_y)$. For simplicity, we assume $W_{xx}=w_{xx}\mathbf{1}_{d_x\times d_x}\in\mathbb{R}^{d_x\times d_x}$, $W_{xy}=w_{xx}\mathbf{1}_{d_x\times d_y}\in\mathbb{R}^{d_x\times d_y}$, $W_{yy}=w_{yy}\mathbf{1}_{d_y\times d_y}\in\mathbb{R}^{d_y\times d_y}$, $\boldsymbol{b}_x=b_x\mathbf{1}_{d_x}\in\mathbb{R}^{d_x}$, and $\boldsymbol{b}_y=b_y\mathbf{1}_{d_y}\in\mathbb{R}^{d_y}$. The following results can be readily extended to matrices W_{xx}, W_{yy}, and W_{xy}, which have a limited number of eigenvalues of $\mathcal{O}(1)$.

In this setting, the analytically continued free energy density $\hat{f}(\beta_{\min},\beta_{\max};\boldsymbol{W})$ calculated using Eq. (6) coincides with the exact free energy density $f(\beta_{\min},\beta_{\max};\boldsymbol{W})$ from Eq. (5).

Theorem 3.1 *For any $\beta_{\min},\beta_{\max}\in\mathbb{R}$ and $w_{xx},w_{xy},w_{yy},b_x,b_y\in\mathbb{R}$, the following equality holds:*

$$f(\beta_{\min},\beta_{\max};\boldsymbol{W})=\hat{f}(\beta_{\min},\beta_{\max};\boldsymbol{W})$$

where

$$f(\beta_{\min},\beta_{\max};\boldsymbol{W})=\operatorname*{extr}_{m^x,m^y}\Bigg[\frac{w_{xx}}{2}(m^x)^2+\frac{\kappa w_{yy}}{2}(m^y)^2+w_{xy}\kappa^{1/2}m^xm^y \\ +b_xm^x+\kappa b_ym^y-\frac{1}{\beta_{\min}}H(m^x)+\frac{\kappa}{\beta_{\max}}H(m^y)\Bigg]$$

where $\kappa={}^{d_y}/_{d_x}$, $H(x)=-x\log(x)-(1-x)\log(1-x)$ denotes binary cross entropy, and extr *denotes the extremum operation.*

This theorem establishes the validity of the analytic continuation for the replica parameter p using Eq. (7) for bilinear min-max games. The detailed proof of this theorem is provided in Appendix. A.

4 Generative Adversarial Networks

Generative adversarial networks (GANs) (Goodfellow et al., 2020) aim to model high-dimensional probability distributions based on training datasets. Despite significant progress in practical applications (Arjovsky et al., 2017; Lucic et al., 2018; Ledig et al., 2017; Isola et al., 2017; Reed et al., 2016), several issues are yet to be resolved, including how the amount of training data influences generalization performance and how sensitive GANs are to specific hyperparameters. This section analyzes the relationship between the amount of training data and generalization error. Additionally, we conduct a sensitivity analysis on the ratio of fake data generated by the generator to the amount of training data, which is critical for the training of GANs. Our analysis employs a minimal setup that captures the intrinsic structure and learning dynamics of GANs (Wang et al., 2019). We consider the high-dimensional limit, where the number of real and fake samples, n and $\tilde{n}$, and the dimension d are large while remaining comparable. Specifically, we analyze the regime in which $n,\tilde{n},d\to\infty$ while maintaining a comparable ratio, i.e., $\alpha={}^{n}/_{d}=\Theta(d^0)$ and $\tilde{\alpha}={}^{\tilde{n}}/_{d}=\Theta(d^0)$, commonly referred to as sample complexity.

4.1 Settings

Generative model for the dataset We consider that a training dataset $\mathcal{D}=\{\boldsymbol{x}^\mu\}_{\mu=1}^n$, where each $\boldsymbol{x}^\mu\in\mathbb{R}^d$ is drawn by the following distribution:

$$\boldsymbol{x}^\mu=\frac{1}{\sqrt{d}}\boldsymbol{w}^*c^\mu+\sqrt{\eta}\boldsymbol{n}^\mu, \tag{10}$$

where $\boldsymbol{w}^*\in\mathbb{R}^d$ is a deterministic feature vector, $c^\mu\in\mathbb{R}$ is random scalar drawn from a standard normal distribution $p(c)=\mathcal{N}(0,1)$, $\boldsymbol{n}^\mu$ is a background noise vector whose components are i.i.d from the standard normal distribution $\mathcal{N}(\mathbf{0}_d, I_d)$, and $\eta \in \mathbb{R}$ is a scalar parameter to control the strength of the noise. We also assume that $\|\boldsymbol{w}^*\|^2 = 1$. This generative model, known as the spiked covariance model (Wishart, 1928; Potters & Bouchaud, 2020), has been studied in statistics to analyze the performance of unsupervised learning methods such as PCA (Ipsen & Hansen, 2019; Biehl & Mietzner, 1993; Hoyle & Rattray, 2004), sparse PCA (Lesieur et al., 2015), deterministic autoencoders (Refinetti & Goldt, 2022), and variational autoencoder (Ichikawa & Hukushima, 2024; 2023).

GAN model Following Wang et al. (2019), we assume that the generator has the same linear structure as the dataset generative model described in Eq. (10):

$$g(z; \boldsymbol{w}) = \frac{1}{\sqrt{d}} \boldsymbol{w} z + \sqrt{\tilde{\eta}} \tilde{\boldsymbol{n}}, \tag{11}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is a learnable parameter, $z \in \mathbb{R}$ is latent variable drawn from a standard normal distribution $p(z) = \mathcal{N}(0, 1)$, $\tilde{\boldsymbol{n}}$ is a noise vector whose components are i.i.d from the standard normal distribution $\mathcal{N}(\mathbf{0}_d, I_d)$, and $\tilde{\eta} \in \mathbb{R}$ is a scalar parameter to control the strength of the noise.

We also define the linear discriminator as

$$d(\boldsymbol{x}; \boldsymbol{v}) = f\left(\frac{1}{\sqrt{d}} \boldsymbol{v}^\top \boldsymbol{x}\right), \tag{12}$$

where $\boldsymbol{x}$ is an input vector, which can be either the real data $\boldsymbol{x}^\mu$ from Eq. (10) or the fake one $g(z^{\tilde{\mu}}; \boldsymbol{w})$ from Eq. (11). The vector $\boldsymbol{v} \in \mathbb{R}^d$ is a learnable parameter, and $f : \mathbb{R} \to \mathbb{R}$ can be any function.

Training algorithm The GAN is trained by solving the following min-max optimization problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \max_{\boldsymbol{v} \in \mathbb{R}^d} V(\boldsymbol{w}, \boldsymbol{v}; \mathcal{D}), \tag{13}$$

where

$$V(\boldsymbol{w}, \boldsymbol{v}; \mathcal{D}, \tilde{\mathcal{D}}) = \sum_{\mu=1}^{n} \phi\left(d(\boldsymbol{x}^\mu; \boldsymbol{v})\right) - \sum_{\tilde{\mu}=1}^{\tilde{n}} \tilde{\phi}\left(d(g(z^{\tilde{\mu}}; \boldsymbol{w}); \boldsymbol{v})\right) - \frac{\lambda}{2}\|\boldsymbol{v}\|^2 + \frac{\tilde{\lambda}}{2}\|\boldsymbol{w}\|^2, \tag{14}$$

where $\tilde{\mathcal{D}} = \{z^\mu\}_{\tilde{\mu}=1}^{\tilde{n}}$ is the latent values of the fake data. The last two terms are regularization terms, where λ and $\tilde{\lambda}$ control the regularization strength. As we assumes a linear discriminator, $V(\boldsymbol{w}, \boldsymbol{v}; \mathcal{D}, \tilde{\mathcal{D}})$ can be expressed as

$$V(\boldsymbol{w}, \boldsymbol{v}; \mathcal{D}, \tilde{\mathcal{D}}) = \sum_{\mu=1}^{n} \phi\left(\frac{1}{\sqrt{d}} \boldsymbol{v}^\top \boldsymbol{x}^\mu\right) - \sum_{\tilde{\mu}=1}^{\tilde{n}} \tilde{\phi}\left(\frac{1}{\sqrt{d}} \boldsymbol{v}^\top g(z^{\tilde{\mu}}; \boldsymbol{w})\right) - \frac{\lambda}{2}\|\boldsymbol{v}\|^2 + \frac{\tilde{\lambda}}{2}\|\boldsymbol{w}\|^2, \tag{15}$$

This value function defined in Eq. (14) is a general form that includes various types of GANs. Specifically, when $\phi = \tilde{\phi}$ and $\|\phi\|_L \leq 1$, it represents a Wasserstein GANs (WGANs) (Arjovsky et al., 2017) and, when $\phi(x) = \log \sigma(x)$ and $\tilde{\phi}(x) = -\log(1 - \sigma(x))$ with σ being the sigmoid function, it corresponds to the Vanilla GANs, which minimize the JS-divergence (Goodfellow et al., 2014a).

Generalization error In the ideal case where the generator perfectly learns the underlying true probability distribution, we have $\boldsymbol{w}^* = \boldsymbol{w}$. Therefore, we define the generalization error ε_g as

$$\varepsilon_g(\bar{\boldsymbol{w}}, \boldsymbol{w}^*) = \frac{1}{d} \mathbb{E}_{\mathcal{D}}\left[\|\bar{\boldsymbol{w}} - \boldsymbol{w}^*\|^2\right], \tag{16}$$

where $\bar{\boldsymbol{w}}$ denotes the min-max optimal value in Eq. (13). The generalization error, ε_g, quantifies the accuracy of signal recovery from the training data.

4.2 Replica Calculation

We apply the replica formalism sketched in Section 2 to derive a set of deterministic equations characterizing the typical behavior of GANs.

In this problem setting, the replicated partition function Z_γ in Eq. (7) can be expressed as

$$Z_\gamma = \int \prod_{a=1}^{\gamma} d\boldsymbol{w}^a \prod_{l=1}^{p} d\boldsymbol{v}^{al} \left(\mathbb{E}_{c,\boldsymbol{n}} e^{\beta_{\max} \sum_{al} \phi\left(\frac{1}{d}(\boldsymbol{v}^{al})^\top \boldsymbol{w}^* c + \sqrt{\frac{\eta}{d}}(\boldsymbol{v}^{al})^\top \boldsymbol{n}\right)} \right)^n$$
$$\times \left(\mathbb{E}_z e^{-\beta_{\max} \sum_{al} \tilde{\phi}\left(\frac{1}{d}(\boldsymbol{v}^{al})^\top \boldsymbol{w}^a z + \sqrt{\frac{\tilde{\eta}}{d}}(\boldsymbol{v}^{al})^\top \tilde{\boldsymbol{n}}\right)} \right)^{\tilde{n}} e^{\frac{\beta_{\max}}{2} \sum_{al}\left(\tilde{\lambda}\|\boldsymbol{w}^a\|^2 - \lambda \|\boldsymbol{v}^{al}\|^2\right)}.$$

To take the average over $\boldsymbol{n}$ and $\tilde{\boldsymbol{n}}$, we notice that since $\boldsymbol{n}$ and $\tilde{\boldsymbol{n}}$ follow a multivariate normal distribution $\mathcal{N}(\boldsymbol{0}_d, I_d)$, the quantities $\boldsymbol{u} = ((\boldsymbol{v}^{al})^\top \boldsymbol{n}/\sqrt{d})_{a,l}$ and $\tilde{\boldsymbol{u}} = ((\boldsymbol{v}^{al})^\top \tilde{\boldsymbol{n}}/\sqrt{d})_{a,l}$ also follow a Gaussian multivariate distribution as

$$p(\boldsymbol{u}) = p(\tilde{\boldsymbol{u}}) = \mathcal{N}(\boldsymbol{0}_{\gamma p}, \boldsymbol{Q}),$$

where

$$\boldsymbol{Q} = (Q_{ls}^{ab}) \in \mathbb{R}^{\gamma p \times \gamma p}, \quad Q_{ls}^{ab} = \frac{1}{d}(\boldsymbol{v}^{al})^\top \boldsymbol{v}^{bs}.$$

To conduct further computations, we introduce auxiliary variables through the following identities:

$$1 = \prod_{abls} d \int \delta(dQ_{ls}^{ab} - (\boldsymbol{v}^{al})^\top \boldsymbol{v}^{bs}) dQ_{ls}^{ab} = \prod_{al} d \int \delta(dm_l^a - (\boldsymbol{v}^{al})^\top \boldsymbol{w}^*) dm_l^a = \prod_{al} d \int \delta(db_l^a - (\boldsymbol{v}^{al})^\top \boldsymbol{w}^a) db_l^a.$$

The replicated partition function can then be expressed as

$$Z_\gamma = \int d\boldsymbol{Q} d\boldsymbol{m} d\boldsymbol{b} e^{\beta_{\min} d(\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b}) + \mathcal{T}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b}))},$$

where we define the entropic term $\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b})$ and energetic term $\mathcal{T}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b})$ as follows:

$$\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b}) \triangleq \frac{1}{d\beta_{\min}} \ln \int \prod_{al} d\boldsymbol{w}^a d\boldsymbol{v}^{al} \prod_{abls} d \int \delta(dQ_{ls}^{ab} - (\boldsymbol{v}^{al})^\top \boldsymbol{v}^{bs})$$
$$\times \prod_{al} d \int \delta(dm_l^a - (\boldsymbol{v}^{al})^\top \boldsymbol{w}^*) d \int \delta(db_l^a - (\boldsymbol{v}^{al})^\top \boldsymbol{w}^a) e^{\frac{\beta_{\max}}{2} \sum_{al}(\tilde{\lambda}\|\boldsymbol{w}^a\|^2 - \lambda\|\boldsymbol{v}^{al}\|^2)},$$
$$\mathcal{T}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b}) \triangleq \frac{\alpha}{\beta_{\min}} \ln \left(\int Dc \int d\boldsymbol{u} p(\boldsymbol{u}) e^{\beta_{\max} \sum_{al} \phi\left(\frac{1}{d}(\boldsymbol{v}^{al})^\top \boldsymbol{w}^* c + \sqrt{\frac{\eta}{d}}(\boldsymbol{v}^{al})^\top \boldsymbol{n}\right)} \right)$$
$$+ \frac{\tilde{\alpha}}{\beta_{\min}} \ln \left(\int Dz \int d\tilde{\boldsymbol{u}} p(\tilde{\boldsymbol{u}}) e^{-\beta_{\max} \sum_{al} \tilde{\phi}\left(\frac{1}{d}(\boldsymbol{v}^{al})^\top \boldsymbol{w}^a z + \sqrt{\frac{\tilde{\eta}}{d}}(\boldsymbol{v}^{al})^\top \tilde{\boldsymbol{n}}\right)} \right).$$

Using the Fourier representation of the delta function, $\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b})$ is further expressed as

$$\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b}) = \frac{1}{d\beta_{\min}} \log \int d\tilde{\boldsymbol{Q}} d\tilde{\boldsymbol{m}} d\tilde{\boldsymbol{b}} e^{d\left(\frac{1}{2}\mathrm{tr}\tilde{\boldsymbol{Q}}\boldsymbol{Q} - \tilde{\boldsymbol{m}}^\top \boldsymbol{m} - \tilde{\boldsymbol{b}}^\top \boldsymbol{b}\right)}$$
$$\left(\int \prod_{al} dw^a dv^{al} e^{-\frac{1}{2}\sum_{abls} \tilde{Q}_{ls}^{ab} v^{al} v^{bs} + w^* \sum_{al} \tilde{m}_l^a v^{al} + \sum_{al} \tilde{b}_l^a w^a v^{al} + \frac{\beta_{\max}}{2} \sum_{al}\left(\tilde{\lambda}(w^a)^2 - \lambda (v^{al})^2\right)} \right)^d. \quad (17)$$

Replica symmetric ansatz Here, we assume the following symmetric structure:

$$\forall a,b \in [\gamma], \forall l,s \in [p],\ Q_{ls}^{ab} = q + \frac{\Delta}{\beta_{\min}}\delta_{ab} + \frac{\chi}{\beta_{\max}}\delta_{ls}\delta_{ab}, \quad (18)$$

$$\forall a,b \in [\gamma], \forall l,s \in [p],\ \tilde{Q}_{ls}^{ab} = \beta_{\max}\hat{q}\delta_{ls}\delta_{ab} - \frac{\beta_{\max}^2}{\beta_{\min}}\hat{\Delta}\delta_{ab} - \beta_{\max}^2\hat{\chi}, \quad (19)$$

$$\forall a \in [\gamma], l \in [p], \quad m_l^a = m, \quad \tilde{m}_l^a = \beta_{\max}\hat{m}, \quad (20)$$

$$\forall a \in [\gamma], l \in [p], \quad b_l^a = b, \quad \hat{b}_l^a = \beta_{\max}\hat{b}. \quad (21)$$

This replica symmetric (RS) structure restricts the integration of the replicated weight parameters $\{\boldsymbol{w}^a\}$, $\{\boldsymbol{v}^{al}\}$ across the entire $\mathbb{R}^{(d\times\gamma p)}\times\mathbb{R}^{(d\times\gamma p)}$ to a subspace that satisfies the constraints in Eq. (18)-(21). This structure, along with scaling by the maximum and minimum beta values, is similar to the standard one-step replica symmetry breaking (1RSB) (Mézard et al., 1987; Takahashi & Kabashima, 2022).

We now turn to the entropic term $\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b})$. The terms that exclude the integrals with respect to $\{\boldsymbol{v}^{al}\}$ and $\{\boldsymbol{w}^a\}$ can be expressed as

$$\begin{aligned}
&\frac{1}{2}\mathrm{tr}\tilde{\boldsymbol{Q}}\boldsymbol{Q}-\tilde{\boldsymbol{m}}^\top\boldsymbol{m}-\tilde{\boldsymbol{b}}^\top\boldsymbol{b}\\
&=\gamma\beta_{\min}\left(-\frac{1}{2}\left(\hat{q}\left(q+\frac{\Delta}{\beta_{\min}}+\frac{\chi}{\beta_{\max}}\right)-\chi\left(\hat{\chi}+\frac{\hat{\Delta}}{\beta_{\min}}\right)+\hat{\chi}\Delta+\hat{\Delta}q+\frac{\Delta\hat{\Delta}}{\beta_{\min}}\right)+\hat{m}m+\hat{b}b\right).
\end{aligned}\tag{22}$$

The term that includes the integrals with respect to $\{\boldsymbol{v}^{al}\}$ and $\{\boldsymbol{w}^a\}$ can be expressed as

$$\begin{aligned}
&\mathbb{E}_z\int\prod_{al}dw^aD\zeta^adv^{al}e^{-\frac{1}{2}\beta_{\max}(\hat{q}+\lambda)\sum_{al}(v^{al})^2+\beta_{\max}\sum_{al}\left(\sqrt{\frac{\hat{\Delta}}{\beta_{\min}}}\zeta^a+\sqrt{\hat{\chi}}z+w^*\hat{m}+w^a\hat{b}\right)v^{al}-\frac{\tilde{\lambda}\beta_{\min}}{2}\sum_a(w^a)^2},\\
&=\mathbb{E}_z\int\prod_a dw^aD\zeta^a\left(\int dv^ae^{-\frac{1}{2}\beta_{\max}(\hat{q}+\lambda)(v^a)^2+\beta_{\max}\left(\sqrt{\frac{\hat{\Delta}}{\beta_{\min}}}\zeta^a+\sqrt{\hat{\chi}}z+w^*\hat{m}+w^a\hat{b}\right)v^a}\right)^pe^{-\frac{\tilde{\lambda}\beta_{\min}}{2}\sum_a(w^a)^2},\\
&=\mathbb{E}_z\int\prod_a dw^aD\zeta^ae^{-\frac{\tilde{\lambda}\beta_{\min}}{2}\sum_a(w^a)^2-\frac{\beta_{\min}}{2(\hat{q}+\lambda)}\sum_a\left(\sqrt{\frac{\hat{\Delta}}{\beta_{\min}}}\zeta^a+\sqrt{\hat{\chi}}z+w^*\hat{m}+w^a\hat{b}\right)^2},\\
&=\mathbb{E}_z\left(\int dwd\zeta e^{\beta_{\min}\left(-\frac{1}{2}\zeta^2-\frac{\tilde{\lambda}}{2}w^2-\frac{(\sqrt{\hat{\Delta}}\zeta+\sqrt{\hat{\chi}}z+w^*\hat{m}+w\hat{b})^2}{2(\hat{q}+\lambda)}\right)}\right)^\gamma.
\end{aligned}$$

This can be derived using the identity, for any $a\in\mathbb{R}_+$ and any $x\in\mathbb{R}$, $e^{\frac{a}{2}x^2}=\int Dze^{\sqrt{a}zx}$. Summarizing these results, the entropic term can be written as

$$\begin{aligned}
\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b},\tilde{\boldsymbol{Q}},\tilde{\boldsymbol{m}},\tilde{\boldsymbol{b}})=\gamma&\left(-\frac{1}{2}\left(\hat{q}\left(q+\frac{\Delta}{\beta_{\min}}+\frac{\chi}{\beta_{\max}}\right)-\chi\left(\hat{\chi}+\frac{\hat{\Delta}}{\beta_{\min}}\right)+\hat{\chi}\Delta+\hat{\Delta}q+\frac{\Delta\hat{\Delta}}{\beta_{\min}}\right)+\hat{m}m+\hat{b}b\right.\\
&\left.+\frac{1}{\beta_{\min}}\int Dz\log\int dwd\zeta e^{\beta_{\min}\left(-\frac{1}{2}\zeta^2-\frac{\tilde{\lambda}}{2}w^2-\frac{(\sqrt{\hat{\Delta}}\zeta+\sqrt{\hat{\chi}}z+w^*\hat{m}+w\hat{b})^2}{2(\hat{q}+\lambda)}\right)}\right).
\end{aligned}$$

By taking the limit as $\beta_{\max}\to\infty$ followed by $\beta_{\min}\to\infty$, owe obtain

$$\mathcal{S}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b},\tilde{\boldsymbol{Q}},\tilde{\boldsymbol{m}},\tilde{\boldsymbol{b}})=-\frac{\gamma}{2}\left(q(\hat{q}+\hat{\Delta})-(\chi-\Delta)\hat{\chi}-2m\hat{m}-2b\hat{b}+\frac{\tilde{\lambda}(\hat{m}^2+\hat{\chi})}{\hat{b}^2+(\hat{q}+\hat{\Delta}+\lambda)\tilde{\lambda}}\right).\tag{23}$$

We next turn to the energetic term $\mathcal{T}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b})$. Under the RS ansatz, $\boldsymbol{u}$ follows

$$u^{al}=\sqrt{\frac{\chi}{\beta_{\max}}}x^{al}+\sqrt{\frac{\Delta}{\beta_{\min}}}y^a+\sqrt{q}\xi,$$

where $\tilde{x}^{al}$, x^{al}, $\tilde{y}^{al}$, y^{al}, ξ, and $\tilde{\xi}$ follow the standard normal distribution $\mathcal{N}(0,1)$. Then, the energetic term $\mathcal{T}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b})$ can be expand as

$$\mathcal{T}(\boldsymbol{Q},\boldsymbol{m},\boldsymbol{b}) \triangleq \underbrace{\frac{\alpha}{\beta_{\min}} \ln\left(\int Dc \int d\boldsymbol{u} p(\boldsymbol{u}) e^{\beta_{\max}\sum_{al}\phi\left(\frac{1}{d}(\boldsymbol{v}^{al})^\top \boldsymbol{w}^* c+\sqrt{\frac{\eta}{d}}(\boldsymbol{v}^{al})^\top \boldsymbol{n}\right)}\right)}_{(a)} + \underbrace{\frac{\tilde{\alpha}}{\beta_{\min}} \ln\left(\int Dz \int d\tilde{\boldsymbol{u}} p(\tilde{\boldsymbol{u}}) e^{-\beta_{\max}\sum_{al}\tilde{\phi}\left(\frac{1}{d}(\boldsymbol{v}^{al})^\top \boldsymbol{w}^a z+\sqrt{\frac{\tilde{\eta}}{d}}(\boldsymbol{v}^{al})^\top \tilde{\boldsymbol{n}}\right)}\right)}_{(b)}.$$

The term (a) can be simplified as

$$\begin{aligned}
\text{(a)} &= \frac{\alpha}{\beta_{\min}} \ln \mathbb{E}_{c,\xi} \int \prod_{al} Dy^a Dx^{al} e^{\beta_{\max}\sum_{al}\phi\left(mc+\sqrt{\eta}\left(\sqrt{\frac{\chi}{\beta_{\max}}}x^{al}+\sqrt{\frac{\Delta}{\beta_{\min}}}y^a+\sqrt{q}\xi\right)\right)}, \\
&= \frac{\alpha}{\beta_{\min}} \ln \mathbb{E}_{c,\xi} \left(\int Dy \left(\int Dx e^{\beta_{\max}\phi\left(mc+\sqrt{\eta}\left(\sqrt{\frac{\chi}{\beta_{\max}}}x+\sqrt{\frac{\Delta}{\beta_{\min}}}y+\sqrt{q}\xi\right)\right)}\right)^p\right)^\gamma, \\
&= \frac{\alpha}{\beta_{\min}} \gamma \mathbb{E}_{c,\xi} \log \int Dy \left(\int Dx e^{\beta_{\max}\phi\left(mc+\sqrt{\eta}\left(\sqrt{\frac{\chi}{\beta_{\max}}}x+\sqrt{\frac{\Delta}{\beta_{\min}}}y+\sqrt{q}\xi\right)\right)}\right)^p + \mathcal{O}(\gamma^2), \\
&= \frac{\alpha}{\beta_{\min}} \gamma \mathbb{E}_{c,\xi} \log \int dy e^{-\frac{\beta_{\min}}{2}y^2} \left(\int dx e^{-\frac{\beta_{\max}}{2}x^2+\beta_{\max}\phi\left(mc+\sqrt{\eta}\left(\sqrt{\chi}x+\sqrt{\Delta}y+\sqrt{q}\xi\right)\right)}\right)^p + \mathcal{O}(\gamma^2).
\end{aligned}$$

Taking the limit as $\beta_{\max} \to \infty$ followed by $\beta_{\min} \to \infty$, we obtain:

$$(a) = \alpha\gamma \mathbb{E}_{c,\xi} \max_y \left[-\frac{1}{2}y^2 - \max_x \left[-\frac{1}{2}x^2 + \phi\left(mc + \sqrt{\eta}\left(\sqrt{\chi}x + \sqrt{\Delta}y + \sqrt{q}\xi\right)\right)\right]\right].$$

Similarly, the term (b) is also expressed as

$$(b) = \frac{\tilde{\alpha}}{\beta_{\min}\gamma} \mathbb{E}_{z,\tilde{\xi}} \ln \int d\tilde{y} e^{-\frac{\beta_{\min}}{2}y^2} \left(\int d\tilde{x} e^{-\frac{\beta_{\max}}{2}\tilde{x}^2-\beta_{\max}\tilde{\phi}\left(bz+\sqrt{\eta}\left(\sqrt{\chi}\tilde{x}+\sqrt{\Delta}\tilde{y}+\sqrt{q}\tilde{\xi}\right)\right)}\right)^p + \mathcal{O}(\gamma^2).$$

Taking the same limits, we find:

$$(b) = \tilde{\alpha}\gamma \mathbb{E}_{z,\tilde{\xi}} \max_{\tilde{y}} \left[-\frac{1}{2}y^2 - \max_{\tilde{x}} \left[-\frac{1}{2}\tilde{x}^2 - \tilde{\phi}\left(bz + \sqrt{\tilde{\eta}}\left(\sqrt{\chi}\tilde{x} + \sqrt{\Delta}\tilde{y} + \sqrt{q}\tilde{\xi}\right)\right)\right]\right].$$

Putting the entropic term and energetic term together, the free energy density is given by

$$f = \underset{\substack{\hat{q},\hat{\chi},\hat{m},\hat{b} \\ q,\delta,\chi,m,b}}{\text{extr}} \frac{1}{2}\left(q\hat{q} - (\chi-\Delta)\hat{\chi} - 2(m\hat{m}+b\hat{b}) + \frac{\tilde{\lambda}(\hat{m}^2+\hat{\chi})}{\hat{b}^2+(\hat{q}+\lambda)\tilde{\lambda}} - 2(\alpha\Phi(q,\Delta,\chi,m,b)+\tilde{\alpha}\tilde{\Phi}(q,\Delta,\chi,m,b))\right), \tag{24}$$

where

$$\Phi(q,\Delta,\chi,m,b) = \mathbb{E}_{c,\xi} \max_y \left[-\frac{1}{2}y^2 - \max_x \left[-\frac{1}{2}x^2 + \phi\left(mc + \sqrt{\eta}\left(\sqrt{\chi}x + \sqrt{\Delta}y + \sqrt{q}\xi\right)\right)\right]\right], \tag{25}$$

$$\tilde{\Phi}(q,\Delta,\chi,m,b) = \mathbb{E}_{z,\tilde{\xi}} \max_{\tilde{y}} \left[-\frac{1}{2}\tilde{y}^2 - \max_{\tilde{x}} \left[-\frac{1}{2}\tilde{x}^2 - \tilde{\phi}\left(bz + \sqrt{\tilde{\eta}}\left(\sqrt{\chi}\tilde{x} + \sqrt{\Delta}\tilde{y} + \sqrt{q}\tilde{\xi}\right)\right)\right]\right]. \tag{26}$$

Note that the min and max operations are involved in the two-level optimization described in Eqs. (25) and (26).

4.3 Results: Application to Simple GANs

In this subsection, following Wang et al. (2019), we apply the formulation derived above to the simple WGAN to demonstrate its generalization properties and conduct a sensitivity analysis of the ratio r fake to real data.

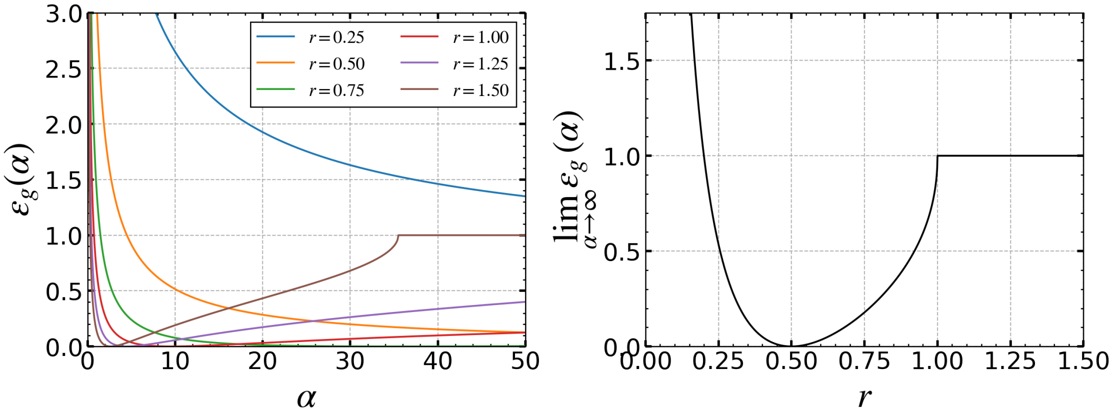

Figure 1: (Left) Generalizatioin error as a function of sample complexity α for different values of the ratio r. (Right) Asymptotic generalization error $\lim_{\alpha\to\infty}\varepsilon(\alpha)$ as a function of the the ratio r.

Self-consistent Equations We consider the case where the functions $\phi(x)$ and $\tilde{\phi}(x)$ are both quadratic, defined as $\phi(x) = \tilde{\phi}(x) = x^2/2$. This setting allows for an explicit calculation of the free energy density, which is given by

$$f = \underset{\substack{q,\chi,m,b\\ \hat{q},\hat{\chi},\hat{m},\hat{b}}}{\operatorname{extr}} \frac{1}{2}\left(q\hat{q} - \chi\hat{\chi} - 2m\hat{m} - 2b\hat{b} + \frac{\tilde{\lambda}(\hat{m}^2+\hat{\chi})}{\hat{b}^2 + (\hat{q}+\lambda)\tilde{\lambda}} - \frac{\alpha(\eta q + m^2)}{\eta\chi - 1} - \frac{\tilde{\alpha}(\tilde{\eta} q + b^2)}{\tilde{\eta}\chi + 1}\right). \tag{27}$$

To find the extremum in Eq. (27), we require that the gradient with respect to each order parameter equals zero. This results in the following set of self-consistent equations:

$$q = \frac{\tilde{\lambda}^2(\hat{m}^2+\hat{\chi})}{(\hat{b}^2+\tilde{\lambda}(\hat{q}+\lambda))^2},\quad \chi = \frac{\tilde{\lambda}}{\hat{b}^2+\tilde{\lambda}(\hat{q}+\lambda)},\quad m = \frac{\hat{m}\tilde{\lambda}}{\hat{b}^2+\tilde{\lambda}(\hat{q}+\lambda)},\quad b = -\frac{\hat{b}\tilde{\lambda}(\hat{m}^2+\hat{\chi})}{(\hat{b}^2+\tilde{\lambda}(\hat{q}+\lambda))^2},$$
$$\hat{q} = \frac{\alpha\eta}{\eta\chi-1} + \frac{\tilde{\alpha}\tilde{\eta}}{\tilde{\eta}\chi+1},\quad \hat{\chi} = \frac{\alpha\eta(q\eta+m^2\rho)}{(\eta\chi-1)^2} + \frac{\tilde{\alpha}\tilde{\eta}(q\tilde{\eta}+d^2\rho)}{(\tilde{\eta}\chi+1)^2},\quad \hat{m} = \frac{\alpha m}{\eta\chi-1},\quad \hat{b} = -\frac{\tilde{\alpha} d}{\tilde{\eta}\chi+1}.$$

Learning Curve For simplicity, we set $\tilde{\alpha} = r\alpha$ and $\lambda = \tilde{\lambda} = \eta = \tilde{\eta} = 1$. Our analysis focuses on how the generalization error depends on α while varying the ratio r, as generating fake data from the generator is generally much easier than collecting real data. Fig. 1 (Left) shows the dependence of generalization error on sample complexity α for various values of the ratio r. The results demonstrate a sharp decline in the generalization error as the ratio r increases. However, when r becomes large, the generalization error increases in the region where α is large, eventually leading to a phase where no learning occurs, and the generalization error equals 1. This implies that as α increases, the learning becomes dominated by only fake data.

In contrast, for smaller r, real data consistently dominates the objective function $V(\boldsymbol{w}, \boldsymbol{v}; \mathcal{D}, \tilde{\mathcal{D}})$, resulting in a steady decrease in generalization error. However, the reduced influence of the fake data component in the objective function, which drives the learning of the generator, requires a significantly larger amount of real data for effective generator training.

Asymptotic Generalisation Error We next analyze the asymptotic behavior of the generalization error when the sample complexity α becomes sufficiently large. The asymptotic behavior of the generalization error as a function of α is given by

$$\varepsilon_g = \begin{cases} \frac{1-2\sqrt{\frac{1-r}{r}}r}{r} + \frac{2\sqrt{2}\left(\sqrt{\frac{1-r}{r}}r+r-1\right)}{(r-1)r\alpha^{1/2}} + \mathcal{O}(\alpha^{-1}) & r \leq 1,\\ 1 + \mathcal{O}(\alpha^{-1}) & r > 1. \end{cases}$$

The results for $\alpha \to \infty$ are shown in Fig. 1(Right). The optimal ratio is $r = 1/2$, indicating that using fake data approximately equal to half of the real data is effective when the dataset approaches infinity. At $r = 1$, a phase transition occurs, suggesting that the model changes from a phase of effective learning phase to one where fake data becomes dominant. Beyond this point, for $r \geq 1$, the model fails to learn any meaningful signal $\boldsymbol{w}^*$, and the generalization error is 1.

Furthermore, when $r = 1/2$, the generalization error scales as $\varepsilon_g \sim \alpha^{-1}$, which represents the optimal asymptotic behavior for a model-matched scenario. These results demonstrate the critical role of the ratio r in determining learning performance. Therefore, tuning the ratio r according to the available amount of real data is crucial for achieving optimal performance. In practice, it is known that in training GANs, the stability of learning can deteriorate depending on the ratio r of fake to real data. This theoretical analysis provides insights into the importance of the ratio r and is expected to contribute to improving learning algorithms.

5 Related Work

The replica method is a non-rigorous but powerful heuristic in statistical physics (Mézard et al., 1987; Mezard & Montanari, 2009; Edwards & Anderson, 1975). It has been proven to be a valuable method for high-dimensional machine-learning problems. Previous studies have investigated the relationship between dataset size and generalization error in supervised learning, including single-layer (Gardner & Derrida, 1988; Opper & Haussler, 1991; Barbier et al., 2019; Aubin et al., 2020) and multi-layer (Aubin et al., 2018) neural networks, as well as kernel methods(Dietrich et al., 1999; Bordelon et al., 2020; Gerace et al., 2020). In unsupervised learning, the replica method has also been applied to dimensionality reduction techniques such as the principal component analysis (Biehl & Mietzner, 1993; Hoyle & Rattray, 2004; 2007), and to generative models such as energy-based models (Decelle et al., 2018; Ichikawa & Hukushima, 2022) and denoising autoencoders (Cui & Zdeborová, 2023). However, the dataset-size dependence of GANs has not been previously analyzed, which this study aims to address. Similar to our work, a statistical mechanical formalism for addressing min-max problems has been proposed (Varga, 1998). However, the treatment of the inverse temperature limit differs from our approach, and it has limitations in accurately handling the order of the min and max operations. In the context of adversarial learning, which involves a non-convex and concave min-max problem, Tanner et al. (2024) analyzes a tractable setting where the internal maximization can be solved. By reducing such cases to standard optimization problems, they apply the replica method and approximate message passing to explore the core phenomenology observed in the adversarial robustness. Even in cases where the internal maximization cannot be explicitly solved, the formalism discussed here provides a basis for further analysis and potential extensions to more complex scenarios.

6 Conclusion

This study introduces a statistical mechanical formalism to analyze high-dimensional min-max optimization problems, focusing on the critical order of min and max operations in non-convex scenarios. Our goal was to perform a sensitivity analysis of equilibrium values, providing new insights into their properties and generalization performance.

We applied this approach to a simple min-max game, evaluated the generalization performance of GANs, and derived the optimal ratio of fake to real data for effective learning. This successful application not only validates the approach but also opens the way for extending this formalism to more complex min-max problems and broader applications, suggesting a promising direction for significant advancements in machine learning and optimization.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

M Biehl and A Mietzner. Statistical mechanics of unsupervised learning. *Europhysics Letters*, 24(5):421, 1993.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Ronald E Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *J. Math. Anal. Appl*, 61(1):159–164, 1977.

Hugo Cui and Lenka Zdeborová. High-dimensional asymptotics of denoising autoencoders. *arXiv preprint arXiv:2305.11041*, 2023.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Thermodynamics of restricted boltzmann machines and related learning dynamics. *Journal of Statistical Physics*, 172:1576–1608, 2018.

Vladimir Fedorovich Dem'yanov and Aleksandr Borisovich Pevnyi. Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12(5):11–52, 1972.

Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical review letters*, 82(14):2975, 1999.

Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

José Fernando Fontanari. A statistical analysis of the knapsack problem. *Journal of Physics A: Mathematical and General*, 28(17):4751, 1995.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1802–1811. PMLR, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014a.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

David C Hoyle and Magnus Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E*, 69(2):026124, 2004.

David C Hoyle and Magnus Rattray. Statistical mechanics of learning multiple orthogonal signals: asymptotic theory and fluctuation effects. *Physical review E*, 75(1):016101, 2007.

Yuma Ichikawa and Koji Hukushima. Statistical-mechanical study of deep boltzmann machine given weight parameters after training by singular value decomposition. *Journal of the Physical Society of Japan*, 91 (11):114001, 2022.

Yuma Ichikawa and Koji Hukushima. Dataset size dependence of rate-distortion curve and threshold of posterior collapse in linear vae. *arXiv preprint arXiv:2309.07663*, 2023.

Yuma Ichikawa and Koji Hukushima. Learning dynamics in linear VAE: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with KL annealing. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1936–1944. PMLR, 02–04 May 2024. URL `https://proceedings.mlr.press/v238/ichikawa24a.html`.

Niels Ipsen and Lars Kai Hansen. Phase transition in pca with missing data: Reduced signal-to-noise ratio, not sample size! In *International Conference on Machine Learning*, pp. 2951–2960. PMLR, 2019.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse pca. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1635–1639. IEEE, 2015.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 907–915. PMLR, 2019.

Pierre-Louis Lions. Une méthode itérative de résolution d’ une inéquation variationnelle. *Israel Journal of Mathematics*, 31:204–208, 1978.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

D Maistroskii. Gradient methods for finding saddle points. *Matekon*, 13(3):22, 1977.

Marc Mezard and Andrea Montanari. *Information, physics, and computation.* Oxford University Press, 2009.

Marc Mézard and Giorgio Parisi. A replica analysis of the travelling salesman problem. *Journal de physique*, 47(8):1285–1296, 1986.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

AS Nemirovsky, DB Yudin, and ER DAWSON. Wiley-interscience series in discrete mathematics, 1983.

Manfred Opper and David Haussler. Generalization performance of bayes optimal classification algorithm for learning a perceptron. *Physical Review Letters*, 66(20):2677, 1991.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists.* Cambridge University Press, 2020.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060–1069. PMLR, 2016.

Maria Refinetti and Sebastian Goldt. The dynamics of representation learning in shallow, non-linear autoencoders. In *International Conference on Machine Learning*, pp. 18499–18519. PMLR, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Takashi Takahashi and Yoshiyuki Kabashima. Macroscopic analysis of vector approximate message passing in a model-mismatched setting. *IEEE Transactions on Information Theory*, 68(8):5579–5600, 2022.

Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, and Florent Krzakala. A high dimensional model for adversarial training: Geometry and trade-offs. *arXiv preprint arXiv:2402.05674*, 2024.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

Peter Varga. Minimax games, spin glasses, and the polynomial-time hierarchy of complexity classes. *Physical Review E*, 57(6):6487, 1998.

John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior.* Princeton university press, 2007.

Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46 (2):265–280, 1945.

Chuang Wang, Hong Hu, and Yue Lu. A solvable high-dimensional model of gan. *Advances in Neural Information Processing Systems*, 32, 2019.

John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pp. 32–52, 1928.

A Derivation of Theorem 3.1 proof

In this section, we provide the derivation proof of Theorem 3.1. The derivation begins with the calculation of the free energy density without the analytic continuation of $p = -\beta_{\min}/\beta_{\max}$ to $p \in \mathbb{N}$. The free energy density in Eq. (5) is connected to the effective Hamiltonian, $\mathcal{H}_{\text{eff}}(\boldsymbol{x}; \boldsymbol{W})$, which is defined through the relationship:

$$f(\beta_{\min}, \beta_{\max}; \boldsymbol{W}) = -\frac{1}{\beta_{\min} d_x} \mathbb{E}_A \log \sum_{\boldsymbol{x}} \exp\left(-\beta_{\min} \mathcal{H}_{\text{eff}}(\boldsymbol{x}; \boldsymbol{W})\right).$$

The effective Hamiltonian is given by

$$\begin{aligned}
\mathcal{H}_{\text{eff}}(\boldsymbol{x}; \boldsymbol{W}) &= \frac{1}{\beta_{\max}} \log \sum_{\boldsymbol{y}} e^{\beta_{\max} V(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{W})}, \\
&= \frac{1}{\beta_{\max}} \log \sum_{\boldsymbol{y}} e^{\beta_{\max}\left(\frac{w_{xx} d_x}{2}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)^2 + \frac{d_y w_{yy}}{2}\left(\frac{\boldsymbol{y}^\top \boldsymbol{1}_{d_y}}{d_y}\right)^2 + w_{xy}\sqrt{d_x d_y}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)\left(\frac{\boldsymbol{y}^\top \boldsymbol{1}_{d_y}}{d_y}\right)\right)} \\
&\times e^{\beta_{\max}\left(b_x d_x \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right) + b_y d_y \left(\frac{\boldsymbol{y}^\top \boldsymbol{1}_{d_y}}{d_y}\right)\right)}, \\
&= \frac{1}{\beta_{\max}} \log e^{\beta_{\max} d_x \left(\frac{w_{xx}}{2}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)^2 + b_x \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)\right)} \\
&\times \sum_{\boldsymbol{y}} e^{\beta_{\max} d_y \left(\frac{w_{yy}}{2}\left(\frac{\boldsymbol{y}^\top \boldsymbol{1}_{d_y}}{d_y}\right)^2 + w_{xy}\sqrt{\frac{d_x}{d_y}}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)\left(\frac{\boldsymbol{y}^\top \boldsymbol{1}_{d_y}}{d_y}\right) + b_y \left(\frac{\boldsymbol{y}^\top \boldsymbol{1}_{d_y}}{d_y}\right)\right)}, \\
&= \beta_{\max} d_x \left(\frac{w_{xx} d_x}{2}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)^2 + b_x d_x \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)\right) \\
&\times \frac{1}{\beta_{\max}} \log \int d\hat{m}^y dm^y e^{\beta_{\max} d_y \left(\frac{w_{yy}}{2}(m^y)^2 + w_{xy}\sqrt{\frac{d_x}{d_y}}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right) m^y + b_y m^y - \frac{1}{\beta_{\max}} m^y \hat{m}^y + \frac{1}{\beta_{\max}} \mathrm{Softplus}(\hat{m}^y)\right) + o(d_y)},
\end{aligned}$$

where $\mathrm{Softplus}(x) = \log(1 + e^x)$. To evaluate the integral with respect to $\hat{m}^y$ and m^y, we take the limit as $d_y \to \infty$ and apply the saddle point approximation. The effective Hamiltonian can be expressed as follows:

$$\begin{aligned}
\mathcal{H}_{\text{eff}}(\boldsymbol{x}; \boldsymbol{W}) = d_x &\left(\frac{w_{xx}}{2}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)^2 + b_x \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right)\right. \\
&\left. + \kappa \operatorname*{extr}_{m^y, \hat{m}^y} \left[\frac{w_{yy}}{2}(m^y)^2 + w_{xy}\kappa^{-1/2}\left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x}\right) m^y + b_y m^y - \frac{1}{\beta_{\max}} m^y \hat{m}^y + \frac{1}{\beta_{\max}} \mathrm{Softplus}(\hat{m}^y)\right]\right).
\end{aligned}$$

where $\kappa = d_y/d_x$. By summing over $\boldsymbol{x}$, the free energy density can be calculated as follows:

$$\begin{aligned}
&f(\beta_{\min}, \beta_{\max}; \boldsymbol{W}) \\
&= -\frac{1}{\beta_{\min} d_x} \log \sum_{\boldsymbol{x}} e^{-\beta_{\min} d_x \left(\frac{w_{xx}}{2} \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x} \right)^2 + b_x \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x} \right) \right)} \\
&\qquad \times e^{-\beta_{\min} d_x \left(\kappa \operatorname*{extr}_{m^y, \hat{m}^y} \left[\frac{w_{yy}}{2} (m^y)^2 + w_{xy} \kappa^{-1/2} \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x} \right) m^y + b_y m^y - \frac{1}{\beta_{\max}} m^y \hat{m}^y + \frac{1}{\beta_{\max}} \mathrm{Softplus}(\hat{m}^y) \right] \right)}, \\
&= -\frac{1}{\beta_{\min} d_x} \log \left(\frac{d_x}{2\pi} \right) \int dm^x d\hat{m}^x e^{-\beta_{\min} d_x \left(\frac{w_{xx}}{2} (m^x)^2 + b_x m^x + \frac{1}{\beta_{\min}} m^x \hat{m}^x - \frac{1}{\beta_{\min}} \mathrm{Softplus}(\hat{m}^x) \right)} \\
&\qquad \times e^{-\beta_{\min} d_x \left(\kappa \operatorname*{extr}_{m^y, \hat{m}^y} \left[\frac{w_{yy}}{2} (m^y)^2 + w_{xy} \kappa^{-1/2} m^x m^y + b_y m^y - \frac{1}{\beta_{\max}} m^y \hat{m}^y + \frac{1}{\beta_{\max}} \mathrm{Softplus}(\hat{m}^y) \right] \right)}, \\
&= \operatorname*{extr}_{m^x, \hat{m}^x, m^y, \hat{m}^y} \left(\frac{w_{xx}}{2} (m^x)^2 + b_x m^x + \frac{1}{\beta_{\min}} m^x \hat{m}^x - \frac{1}{\beta_{\min}} \mathrm{Softplus}(\hat{m}^x) \right. \\
&\qquad \left. + \frac{\kappa w_{yy}}{2} (m^y)^2 + w_{xy} \kappa^{1/2} m^x m^y + \kappa b_y m^y - \frac{\kappa}{\beta_{\max}} m^y \hat{m}^y + \frac{\kappa}{\beta_{\max}} \mathrm{Softplus}(\hat{m}^y) \right).
\end{aligned}$$

The final equality is obtained by applying the saddle point method to evaluate the integral. From the saddle point equations, the following expressions are

$$m^x = \sigma(\hat{m}^x), \quad m^y = \sigma(\hat{m}^y)$$

Further transformation of the equation yields the following expression:

$$f(\beta_{\min}, \beta_{\max}; \boldsymbol{W}) = \operatorname*{extr}_{m^x, m^y} \left[\frac{w_{xx}}{2} (m^x)^2 + \frac{\kappa w_{yy}}{2} (m^y)^2 + w_{xy} \kappa^{1/2} m^x m^y + b_x m^x + \kappa b_y m^y - \frac{1}{\beta_{\min}} H(m^x) + \frac{\kappa}{\beta_{\max}} H(m^y) \right], \tag{28}$$

where $H(x) = -x \log x - (1-x) \log(1-x)$ represents the binary cross-entropy.

Next, we proceed to evaluate the free energy density under analytic continuation in the replica method, which is expressed as

$$\begin{aligned}
\hat{f}(\beta_{\min}, \beta_{\max}; \boldsymbol{W}) = -\frac{1}{\beta_{\min} d_x} \log \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p} & e^{\beta_{\max} d_x \left(\frac{w_{xx} p}{2} \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x} \right)^2 + \frac{\kappa w_{yy}}{2} \sum_l \left(\frac{\boldsymbol{y}_l^\top \boldsymbol{1}_{d_y}}{d_y} \right)^2 \right)} \\
& \times e^{\beta_{\max} d_x \left(w_{xy} \kappa^{1/2} \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x} \right) \sum_l \left(\frac{\boldsymbol{y}_l^\top \boldsymbol{1}_{d_y}}{d_y} \right) + b_x p \left(\frac{\boldsymbol{x}^\top \boldsymbol{1}_{d_x}}{d_x} \right) + \kappa b_y \sum_l \left(\frac{\boldsymbol{y}_l^\top \boldsymbol{1}_{d_y}}{d_y} \right) \right)}.
\end{aligned}$$

We introduce the order parameter through the Fourier transform representation of the delta function:

$$\begin{aligned}
&\hat{f}(\beta_{\min}, \beta_{\max}; \boldsymbol{W}) \\
&= -\frac{1}{\beta_{\min} d_x} \log \left(\frac{d_x d_y^p}{(2\pi)^{p+1}} \right) \int dm^x d\hat{m}^x \prod_l dm_l^y d\hat{m}_l^y e^{\beta_{\max} d_x \left(\frac{w_{xx} p}{2} (m^x)^2 + \frac{\kappa w_{yy}}{2} \sum_l (m_l^y)^2 + w_{xy} \kappa^{1/2} m^x \sum_l m_l^y \right)} \\
&\times e^{\beta_{\max} d_x \left(b_x p m^x + \kappa b_y \sum_l m_l^y - \frac{1}{\beta_{\max}} \hat{m}^x m^x - \frac{\kappa}{\beta_{\max}} \sum_l \hat{m}_l^y m_l^y \right)} \sum_{\boldsymbol{x}} e^{\hat{m}^x \sum_i x_i} \sum_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p} e^{\sum_l \hat{m}_l^y \sum_j y_j}, \\
&= -\frac{1}{\beta_{\min} d_x} \log \left(\frac{d_x d_y^p}{(2\pi)^{p+1}} \right) \int dm^x d\hat{m}^x \prod_l dm_l^y d\hat{m}_l^y e^{\beta_{\max} d_x \left(\frac{w_{xx} p}{2} (m^x)^2 + \frac{\kappa w_{yy}}{2} \sum_l (m_l^y)^2 + w_{xy} \kappa^{1/2} m^x \sum_l m_l^y \right)} \\
&\times e^{\beta_{\max} d_x \left(b_x p m^x + \kappa b_y \sum_l m_l^y - \frac{1}{\beta_{\max}} \hat{m}^x m^x - \frac{\kappa}{\beta_{\max}} \sum_l \hat{m}_l^y m_l^y + \frac{1}{\beta_{\max}} \mathrm{Softplus}(\hat{m}^x) + \frac{\kappa}{\beta_{\max}} \sum_l \mathrm{Softplus}(\hat{m}_l^y) \right)}.
\end{aligned}$$

Under the assumption of replica symmetry, where $\forall l \in [p], \quad \hat{m}_l^y = m^y, m_l^y = m^y$, we can further reformulate the expression as follows:

$$\begin{aligned}
&\hat{f}(\beta_{\min}, \beta_{\max}; \boldsymbol{W}) \\
&= -\frac{1}{\beta_{\min} d_x} \log \int dm^x d\hat{m}^x dm^y d\hat{m}^y e^{-\beta_{\min} d_x \left(\frac{w_{xx}}{2}(m^x)^2 + \frac{\kappa w_{yy}}{2}(m^y)^2 + w_{xy}\kappa^{1/2} m^x m^y \right)} \\
&\qquad \times e^{-\beta_{\min} d_x \left(b_x m^x + \kappa b_y m^y - \frac{1}{\beta_{\max} p}\hat{m}^x m^x - \frac{\kappa}{\beta_{\max}}\hat{m}^y m^y + \frac{1}{\beta_{\max} p}\mathrm{Softplus}(\hat{m}^x) + \frac{\kappa}{\beta_{\max}}\mathrm{Softplus}(\hat{m}^y) + o(d^x) + o(d^y) \right)}, \\
&= \underset{m^x, m^y, \hat{m}^x, \hat{m}^y}{\mathrm{extr}} \left[\frac{w_{xx}}{2}(m^x)^2 + \frac{\kappa w_{yy}}{2}(m^y)^2 + w_{xy}\kappa^{1/2} m^x m^y + b_x m^x + \kappa b_y m^y \right. \\
&\qquad \left. - \frac{1}{\beta_{\max} p}\hat{m}^x m^x - \frac{\kappa}{\beta_{\max}}\hat{m}^y m^y + \frac{1}{\beta_{\max} p}\mathrm{Softplus}(\hat{m}^x) + \frac{\kappa}{\beta_{\max}}\mathrm{Softplus}(\hat{m}^y) \right].
\end{aligned}$$

The final equality is derived by handling the integral using the saddle point method. Consequently, the following saddle point equation is obtained:

$$m^x = \sigma(\hat{m}^x), \quad m^y = \sigma(\hat{m}^y).$$

Substituting these results, the following expression for the free energy density is derived as

$$\hat{f}(\beta_{\min}, \beta_{\max}; \boldsymbol{W}) = \underset{m^x, m^y}{\mathrm{extr}} \left[\frac{w_{xx}}{2}(m^x)^2 + \frac{\kappa w_{yy}}{2}(m^y)^2 + w_{xy}\kappa^{1/2} m^x m^y + b_x m^x + \kappa b_y m^y - \frac{1}{\beta_{\min}} H(m^x) + \frac{\kappa}{\beta_{\max}} H(m^y) \right]. \tag{29}$$

This result coincides with the exact free energy density $f(\beta_{\min}, \beta_{\max}; \boldsymbol{W})$, derived without the need for analytic continuation.